Case Study 2 - Analyzing data from MovieLens

Data Science with R

Introduction

In this case study we will look at the movies data set from MovieLens. It contains data about users and how they rate movies.

Problem 1: Importing the MovieLens data set and merging it into a single data frame

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
mlData <- distinct(mlData)
print(colnames(mlData))
```

```
## [1] "user_id"      "movie_title"  "genre"        "rating"       "release_date"
## [6] "age"          "gender"       "occupation"
```

Report some basic details of the data you collected. For example:

- How many movies have an average rating over 4.5 overall?
  - I found there were 11 movies with a mean rating over 4.5.

```
mlData_aggregates <- mlData %>%
  group_by(movie_title) %>%
  summarise(mean_rating = mean(rating, na.rm = TRUE))

high_mean_rating_count <- mlData_aggregates %>%
  filter(mean_rating > 4.5) %>%
  nrow()

mlData_aggregates %>%
  filter(mean_rating > 4.5) %>%
  head(high_mean_rating_count)
```

```
## # A tibble: 11 x 2
##    movie_title                                          mean_rating
##    <fct>                                                      <dbl>
##  1 "Aiqing wansui (1994)"                                         5
```

```
##  2 "Entertaining Angels: The Dorothy Day Story (1996)"        5
##  3 "Great Day in Harlem, A (1994)"                            5
##  4 "Marlene Dietrich: Shadow and Light (1996) "               5
##  5 "Pather Panchali (1955)"                                   4.62
##  6 "Prefontaine (1997)"                                       5
##  7 "Saint of Fort Washington, The (1993)"                     5
##  8 "Santa with Muscles (1996)"                                5
##  9 "Someone Else's America (1995)"                            5
## 10 "Star Kid (1997)"                                          5
## 11 "They Made Me a Criminal (1939)"                           5
```

- How many movies have an average rating over 4.5 among men?
  - I found there were 18 movies with a mean rating over 4.5 among women.

```
mlData_aggregates <- mlData %>%
  group_by(movie_title) %>%
  filter(gender == "M") %>%
  summarise(mean_rating_men = mean(rating, na.rm = TRUE)) %>%
  full_join(mlData_aggregates)
```

```
## Joining, by = "movie_title"
```

```
high_mean_men_count <- mlData_aggregates %>%
  filter(mean_rating_men > 4.5) %>%
  nrow()

mlData_aggregates %>%
  arrange(desc(mean_rating_men)) %>%
  select(-mean_rating) %>%
  head()
```

```
## # A tibble: 6 x 2
##   movie_title                                          mean_rating_men
##   <fct>                                                          <dbl>
## 1 Aiqing wansui (1994)                                               5
## 2 Delta of Venus (1994)                                              5
## 3 Entertaining Angels: The Dorothy Day Story (1996)                  5
## 4 Great Day in Harlem, A (1994)                                      5
## 5 Leading Man, The (1996)                                            5
## 6 Letter From Death Row, A (1998)                                    5
```

- How many movies have an average rating over 4.5 among women?
  - I found by using a similar approach to the above that there were 16 movies with a mean rating over 4.5 among women.

```
## Joining, by = "movie_title"
```

```
## # A tibble: 6 x 2
##   movie_title                              mean_rating_women
##   <fct>                                                <dbl>
## 1 Everest (1998)                                           5
## 2 Faster Pussycat! Kill! Kill! (1965)                      5
## 3 Foreign Correspondent (1940)                             5
## 4 Maya Lin: A Strong Clear Vision (1994)                   5
## 5 Mina Tannenbaum (1994)                                   5
## 6 Prefontaine (1997)                                       5
```

- Let us order by mean rating but keep men/women mean rating columns for comparison. Note that some movies were not rated by both men and women.

```
## # A tibble: 10 x 4
##    movie_title                       mean_rating_wom~ mean_rating_men mean_rating
##    <fct>                                        <dbl>           <dbl>       <dbl>
##  1 "Prefontaine (1997)"                             5               5           5
##  2 "Someone Else's America (1995)"                  5              NA           5
##  3 "Aiqing wansui (1994)"                          NA               5           5
##  4 "Entertaining Angels: The Dorot~                NA               5           5
##  5 "Great Day in Harlem, A (1994)"                 NA               5           5
##  6 "Marlene Dietrich: Shadow and L~                NA               5           5
##  7 "Saint of Fort Washington, The ~                NA               5           5
##  8 "Santa with Muscles (1996)"                     NA               5           5
##  9 "Star Kid (1997)"                               NA               5           5
## 10 "They Made Me a Criminal (1939)"                NA               5           5
```

- How many movies have an median rating over 4.5 among men over age 30?
  - I found there were 47 movies with a median rating over 4.5 among men over 30.

```
mlData_aggregates <- mlData %>%
  group_by(movie_title) %>%
  filter(gender == "M", age > 30) %>%
  summarise(median_rating_men30plus = median(rating)) %>%
  full_join(mlData_aggregates)
```

```
## Joining, by = "movie_title"
```

```
high_median_men30plus_count <- mlData_aggregates %>%
  filter(median_rating_men30plus > 4.5) %>%
  nrow()

mlData_aggregates %>%
  arrange(desc(median_rating_men30plus)) %>%
  select(movie_title, median_rating_men30plus) %>%
  head(10)
```

```
## # A tibble: 10 x 2
##    movie_title                                          median_rating_men30plus
##    <fct>                                                                 <dbl>
##  1 Aiqing wansui (1994)                                                      5
##  2 Anna (1996)                                                               5
##  3 Aparajito (1956)                                                          5
##  4 Big Sleep, The (1946)                                                     5
##  5 Casablanca (1942)                                                         5
##  6 Citizen Kane (1941)                                                       5
##  7 Close Shave, A (1995)                                                     5
##  8 Delta of Venus (1994)                                                     5
##  9 Entertaining Angels: The Dorothy Day Story (1996)                         5
## 10 Faithful (1996)                                                           5
```

-How many movies have an median rating over 4.5 among women over age 30?

- I found using a simlar approach to the above that there were 70 movies with a median rating over 4.5 among women over 30.

```
## Joining, by = "movie_title"
```

```
## [1] 70
```

```
## # A tibble: 10 x 2
##    movie_title              median_rating_women30plus
##    <fct>                                        <dbl>
##  1 Amateur (1994)                                   5
##  2 Angel Baby (1995)                                5
##  3 Bent (1997)                                      5
##  4 Best Men (1997)                                  5
##  5 Big Lebowski, The (1998)                         5
##  6 Blade Runner (1982)                              5
##  7 Brassed Off (1996)                               5
##  8 Braveheart (1995)                                5
##  9 Casablanca (1942)                                5
## 10 Cats Don't Dance (1997)                          5
```

- For comparison, order by median rating but keep men/women over 30 median rating columns. Note here that some movies were not rated by both men over 30 and women over 30.

```
## Joining, by = "movie_title"
```

```
## # A tibble: 10 x 4
##    movie_title              median_rating median_rating_men~ median_rating_wome~
##    <fct>                            <dbl>              <dbl>                <dbl>
##  1 Aiqing wansui (1994)                 5                  5                   NA
##  2 Aparajito (1956)                     5                  5                   NA
##  3 Casablanca (1942)                    5                  5                    5
##  4 Citizen Kane (1941)                  5                  5                    4
##  5 Close Shave, A (1995)                5                  5                    5
##  6 Entertaining Angels: Th~             5                  5                   NA
##  7 Faust (1994)                         5                  5                   NA
##  8 Godfather, The (1972)                5                  5                    4
##  9 Great Day in Harlem, A ~             5                  5                   NA
## 10 Hugo Pool (1997)                     5                  1                   NA
```

- What are the ten most "popular" movies?
  - Perhaps we might consider a movie popular if it has both a high mean and median rating. I found there were many films with a mean rating of 5 and many films with a median rating of 5. Upon finding the intersection of these two sets, it turned out that there were 10 films with both a mean rating of 5 and a median rating of 5. Without considering rating count, we could propose that the top ten most popular films are these 10 films with median and mean rating of 5. ‘

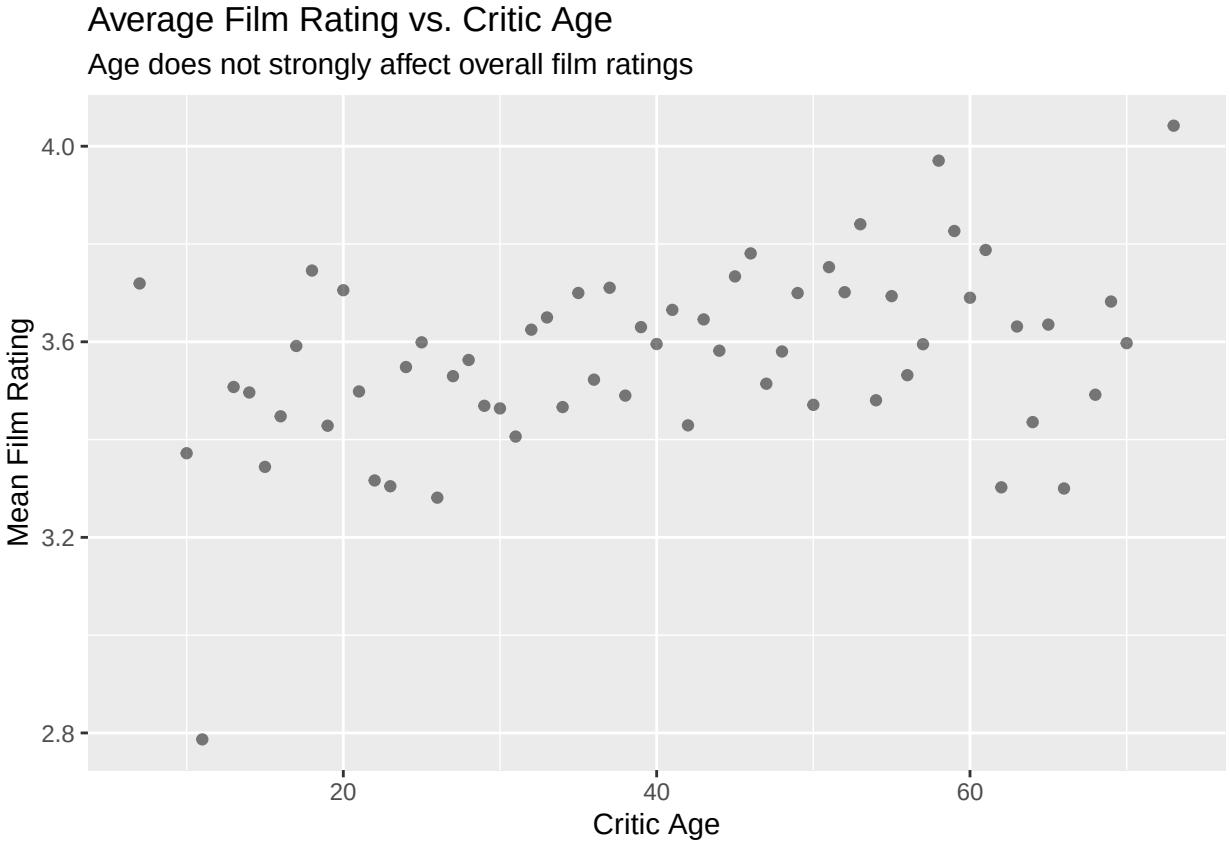
```
mlData_popular <- mlData_aggregates %>%
  filter(
    (mean_rating == 5) & (median_rating == 5)
    )

mlData_popular %>%
  select(movie_title, median_rating, mean_rating) %>%
  head(nrow(mlData_popular))
```

```
## # A tibble: 10 x 3
##    movie_title                                          median_rating mean_rating
##    <fct>                                                        <dbl>       <dbl>
##  1 "Aiqing wansui (1994)"                                           5           5
##  2 "Entertaining Angels: The Dorothy Day Story (1996)"              5           5
##  3 "Great Day in Harlem, A (1994)"                                  5           5
```

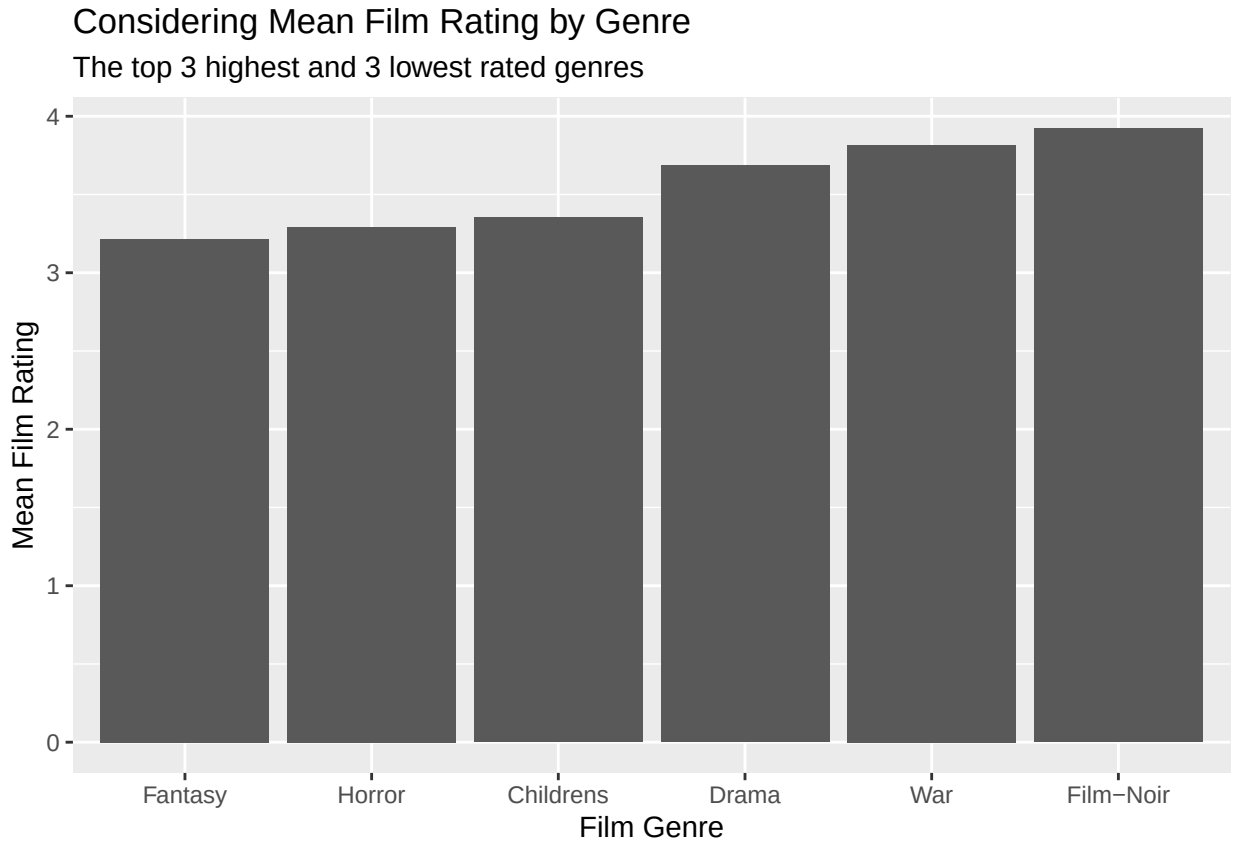
```
##  4 "Marlene Dietrich: Shadow and Light (1996) "                    5           5
##  5 "Prefontaine (1997)"                                            5           5
##  6 "Saint of Fort Washington, The (1993)"                          5           5
##  7 "Santa with Muscles (1996)"                                     5           5
##  8 "Someone Else's America (1995)"                                 5           5
##  9 "Star Kid (1997)"                                               5           5
## 10 "They Made Me a Criminal (1939)"                                5           5
```

- Make some conjectures about how easy various groups are to please!
  - Question: Does the mean rating of all films depend on the age of the reviewer?
  - Answer: It appears not, at least without grouping by further characteristics.

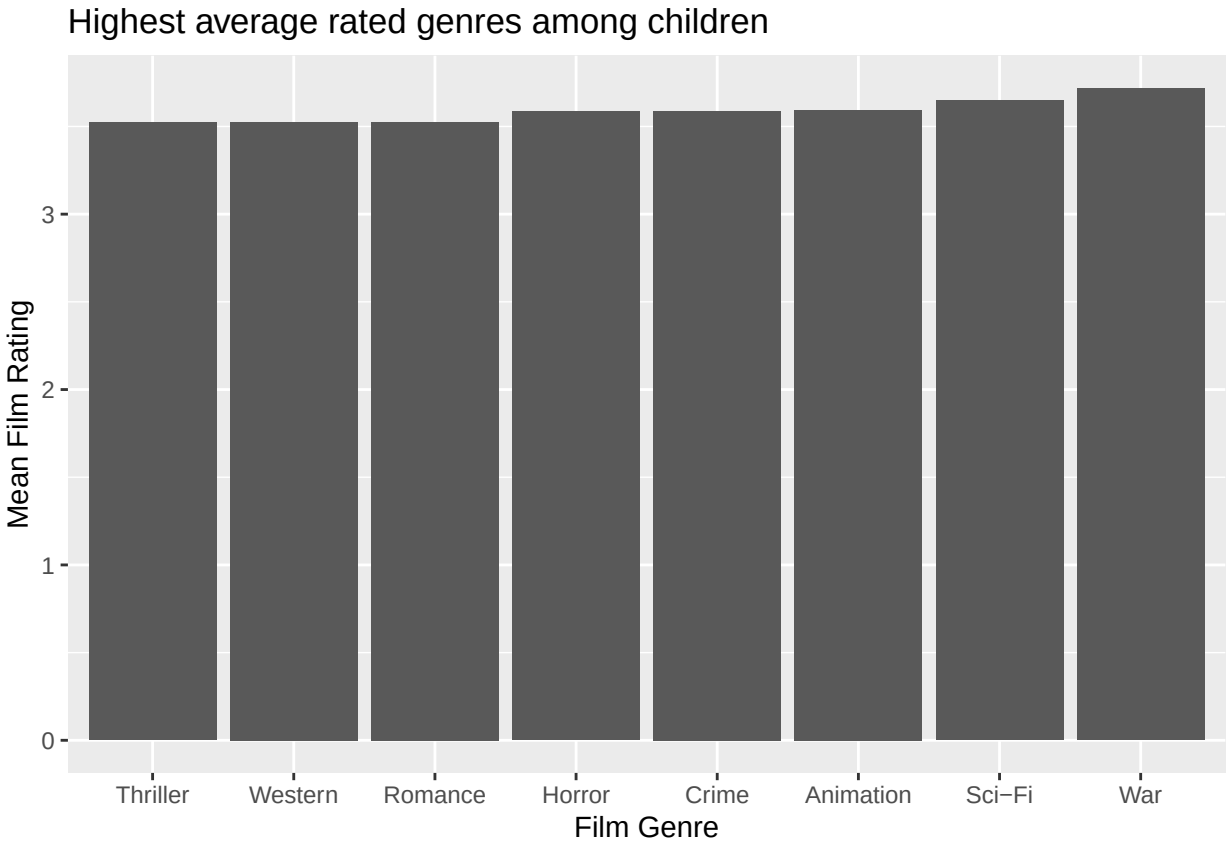


- Question: Do some film genres just generally receive higher ratings than other genres? Do some film genres perform well with certain groups but poorly with other groups?
- Answer: The highest rated genre (using mean rating of all films within each genre) is Film-Noir with a mean rating of about 3.92 while the lowest rated genre is Fantasy with a mean rating of about 3.22.
- Answer: Usually there is not much difference between ratings by men vs women but men tend to enjoy Film-Noir more than women while women enjoy Musicals more than men.
- Answer: Children (critics aged less than 16) enjoy war, sci-fi, and animation films the most.

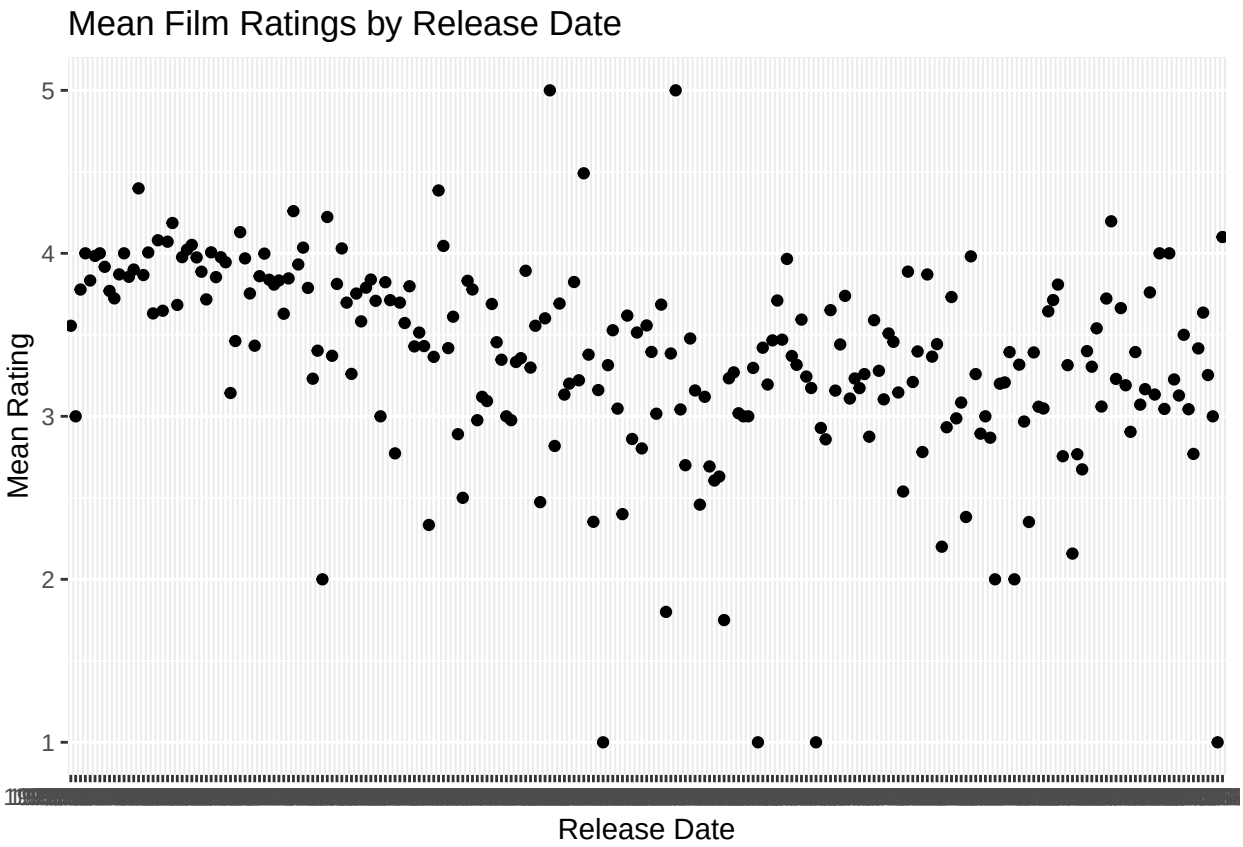
```
## Joining, by = "genre"
## Joining, by = "genre"
```



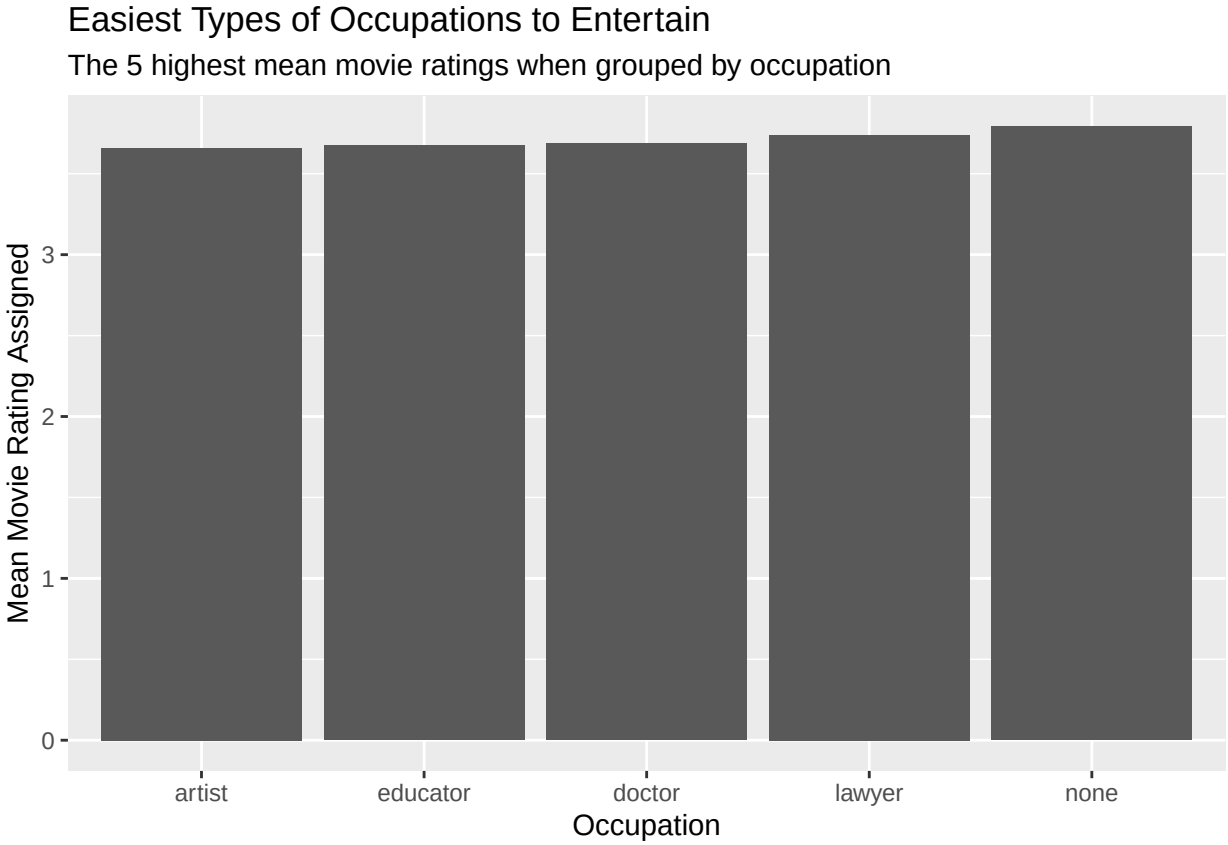

```
## # A tibble: 5 x 3
##   genre     mean_rating_men mean_rating_women
##   <fct>               <dbl>             <dbl>
## 1 Film-Noir            3.97              3.74
## 2 Mystery              3.67              3.56
## 3 Western              3.64              3.51
## 4 Musical              3.47              3.64
## 5 Childrens            3.32              3.43
```

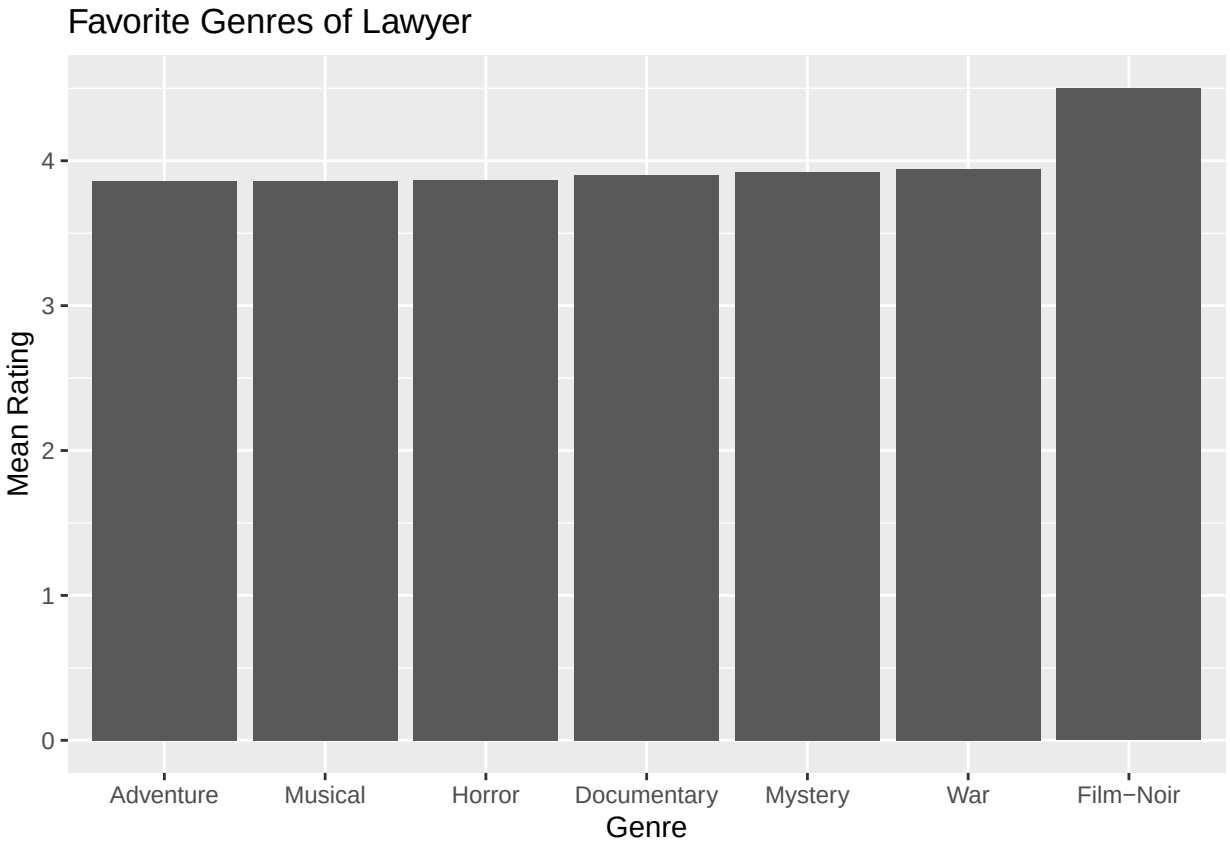


+ Question: How do critic ratings depend on how old a film is? + Answer: Apparently no, but there appears to be more variance in ratings among newer films.



- Question: How does occupation relate to mean ratings?
- Answer: The occupations that tend to rate movies the highest are the unemployed, doctors, lawyers, educators, and artists. Since the unemployed are the easiest to please we might consider focusing on this group. However, the unemployed may not have as much money to spend on movies, so consider next what types of films are best liked by lawyers.

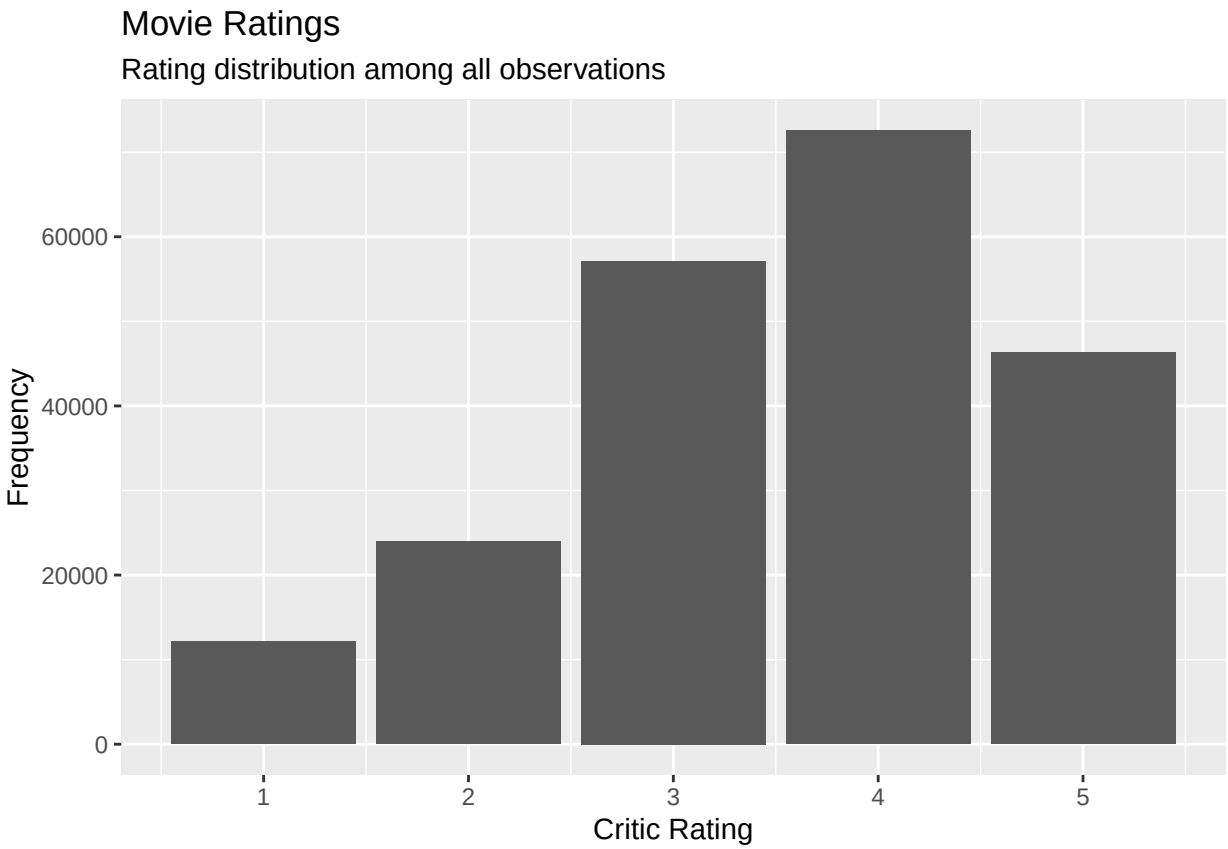

Easiest Types of Occupations to Entertain
The 5 highest mean movie ratings when grouped by occupation
Mean Movie Rating Assigned
0
1
2
3
artist
educator
doctor
lawyer
none
Occupation



Problem 2: Expand our investigation to histograms

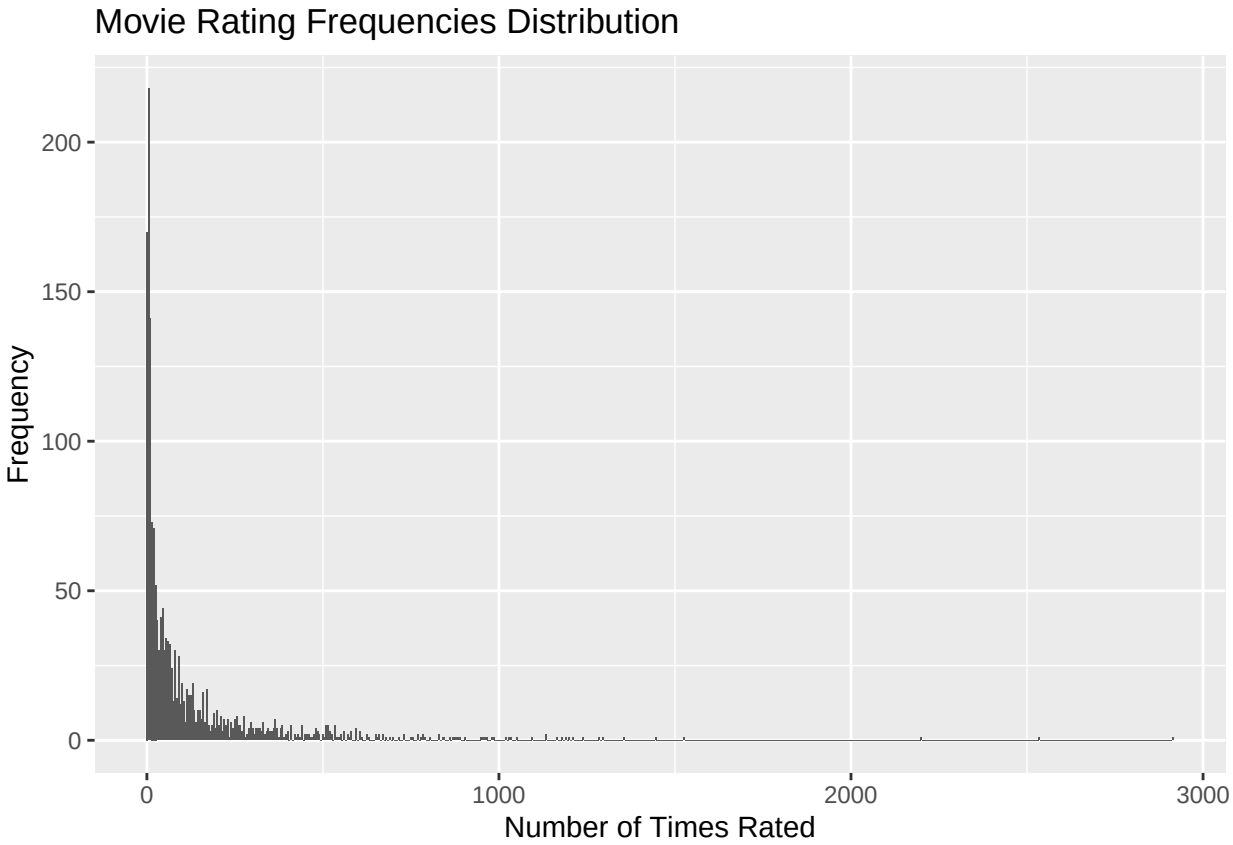
An obvious issue with any inferences drawn from Problem 1 is that we did not consider how many times a movie was rated.

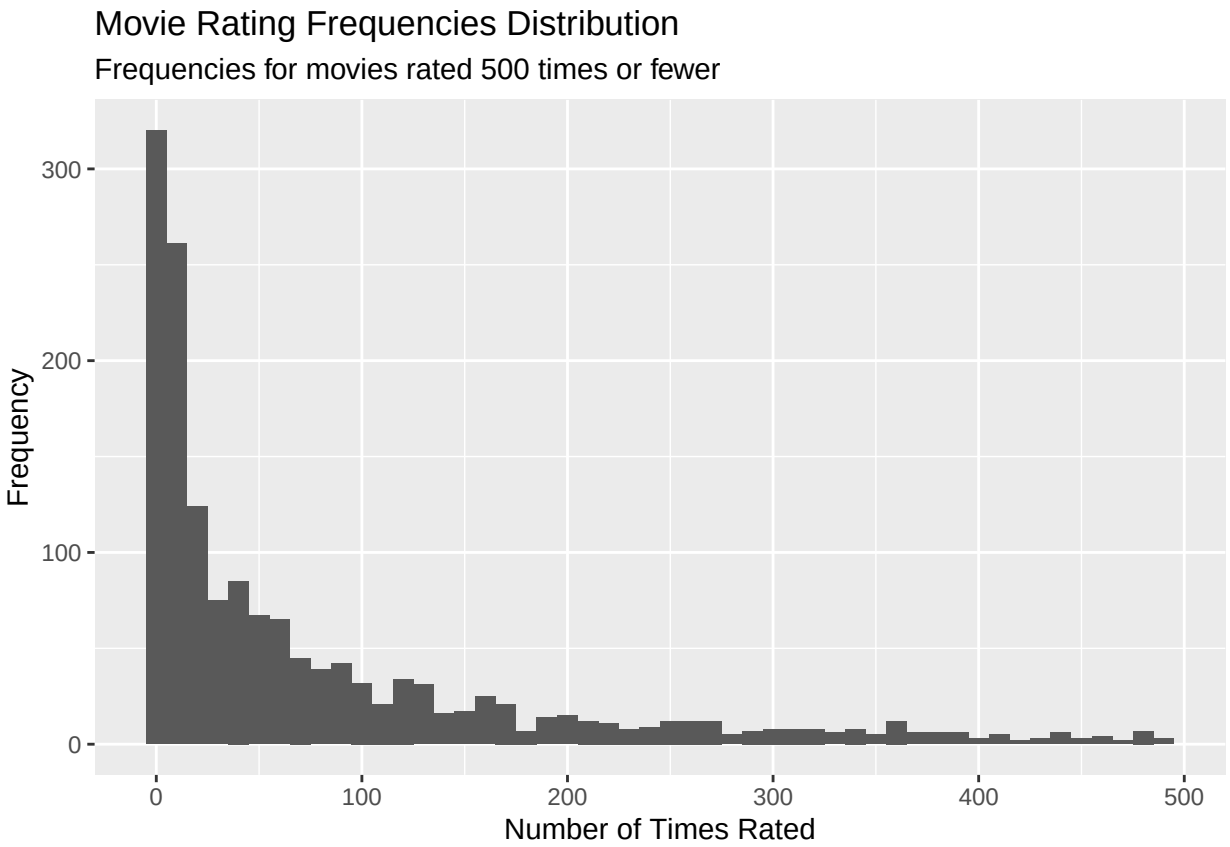
- Plot a histogram of the ratings of all movies.



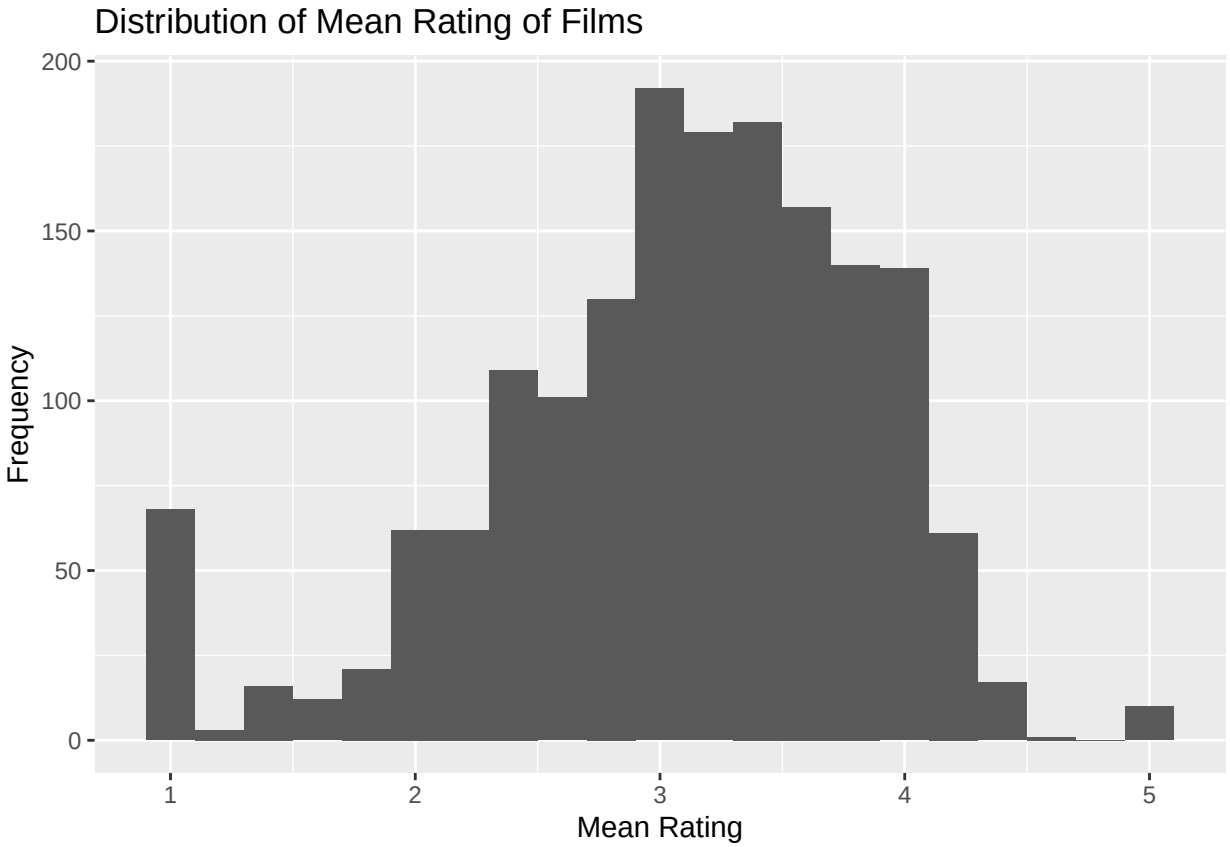
- Plot a histogram of the number of ratings each movie received.

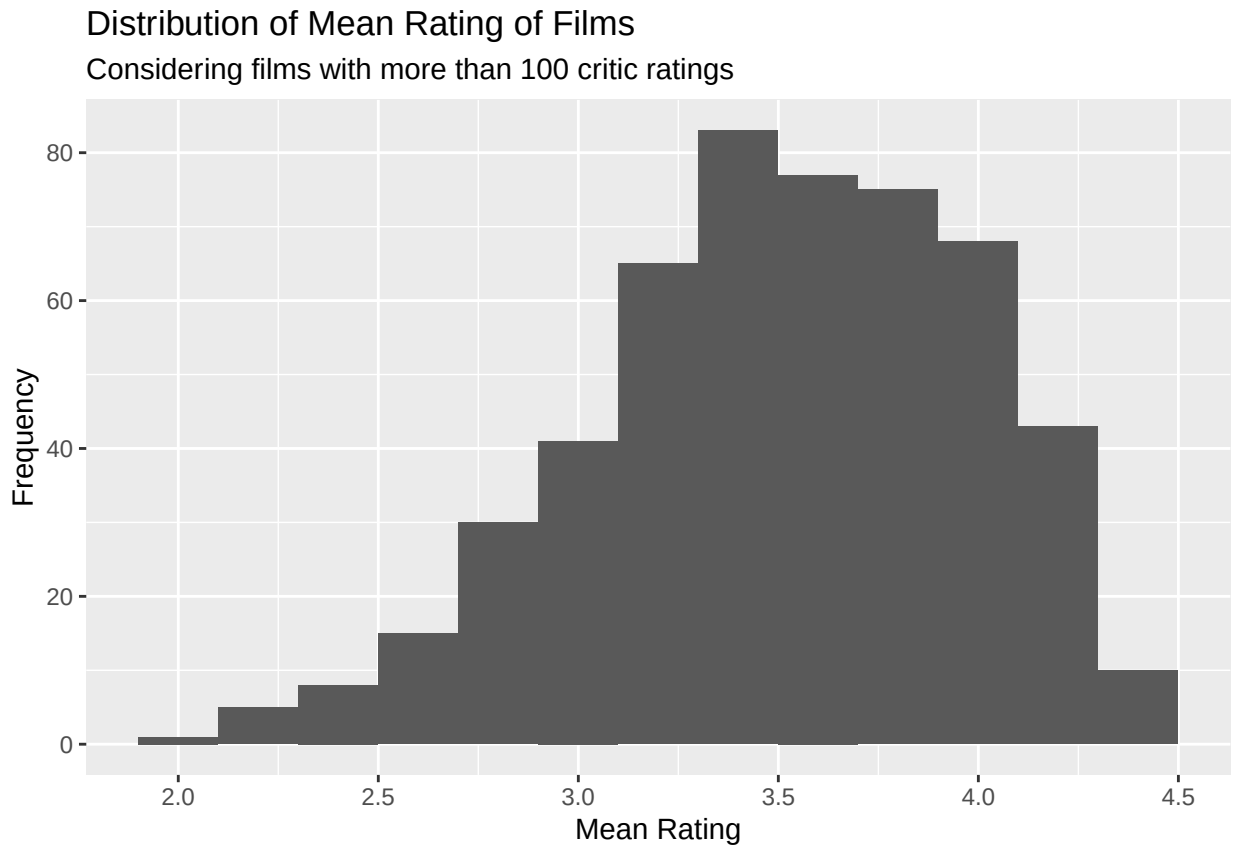
```
## Joining, by = "movie_title"
```


Movie Rating Frequencies Distribution
Frequency
200
150
100
50
0
0
1000
2000
3000
Number of Times Rated



- Plot a histogram of the average rating for each movie.
- Plot a histogram of the average rating for movies which are rated more than 100 times.
  - Notice that when we include movies with 100 or fewer ratings, there are more mean ratings on the ends of the distribution. So when we reduce the dataset to just films with more than 100 ratings, the distribution of rating means tends to have lower variance.
  - Generally speaking, it is better to trust that a movie with high mean rating and a high number (>100) of critic ratings than a movie with high mean rating and a low number (<= 100) of critic ratings. The reason is that infrequently rated movies are more likely to have very high or very low mean. In contrast, frequently rated films are more likely to have a moderate mean rating (between 2 and 4). So a frequently rated film with high mean rating is robust to increases in the number of ratings while an infrequently rated film with high mean may have a high mean due to chance.


Distribution of Mean Rating of Films
Frequency
200
150
100
50
0
1
2
3
4
5
Mean Rating



- Make some conjectures about the distribution of ratings!
  - Question: We saw that movies with a large number of ratings or few ratings may tend to have more extreme results. Do films with a large number of ratings do better or worse than those with a moderate number of ratings? What about films with very few ratings.
  - Answer: It looks like films that are rated often tend to have a higher mean rating compared to the whole dataset while films that have very few ratings have a lower mean rating compared to the whole dataset.

```
count_deciles = quantile(mlData_aggregates$rating_count, c(.1, .2, .3, .4, .5, .6, .7, .8, .9))
count_deciles
```

```
##   10%   20%   30%   40%   50%   60%   70%   80%   90%
##   2.0   6.0  12.0  22.0  42.0  65.6 108.0 188.4 363.9
```

```
oft_rated_films_df <- mlData %>%
  group_by(movie_title) %>%
  summarise(
    rating_count = n(),
    median_rating = median(rating),
    mean_rating = mean(rating)
    ) %>%
  filter(rating_count > count_deciles[9])
head(oft_rated_films_df)
```

```
## # A tibble: 6 x 4
##   movie_title                  rating_count median_rating mean_rating
##   <fct>                               <int>         <dbl>       <dbl>
```

```
## 1 2001: A Space Odyssey (1968)         1036             4        3.97
## 2 Abyss, The (1989)                     604             4        3.59
## 3 African Queen, The (1951)             608             4        4.18
## 4 Air Force One (1997)                  862             4        3.63
## 5 Aladdin (1992)                        876             4        3.81
## 6 Alien (1979)                         1164             4        4.03
```

```
rarely_rated_films_df <- mlData %>%
  group_by(movie_title) %>%
  summarise(
    rating_count = n(),
    median_rating = median(rating),
    mean_rating = mean(rating)
    ) %>%
  filter(rating_count < count_deciles[3])
head(rarely_rated_films_df)
```

```
## # A tibble: 6 x 4
##   movie_title                              rating_count median_rating mean_rating
##   <fct>                                           <int>         <dbl>       <dbl>
## 1 1-900 (1994)                                        5             3        2.6
## 2 3 Ninjas: High Noon At Mega Mountain (~            10             1        1
## 3 8 Heads in a Duffel Bag (1997)                      4             4        3.25
## 4 8 Seconds (1994)                                    4             4        3.75
## 5 A Chef in Love (1996)                               8             4        4.12
## 6 Á köldum klaka (Cold Fever) (1994)                  2             3        3
```

```
mean(oft_rated_films_df$mean_rating)
```

```
## [1] 3.631668
```

```
mean(rarely_rated_films_df$mean_rating)
```

```
## [1] 2.621521
```

```
mean(mlData_aggregates$mean_rating)
```

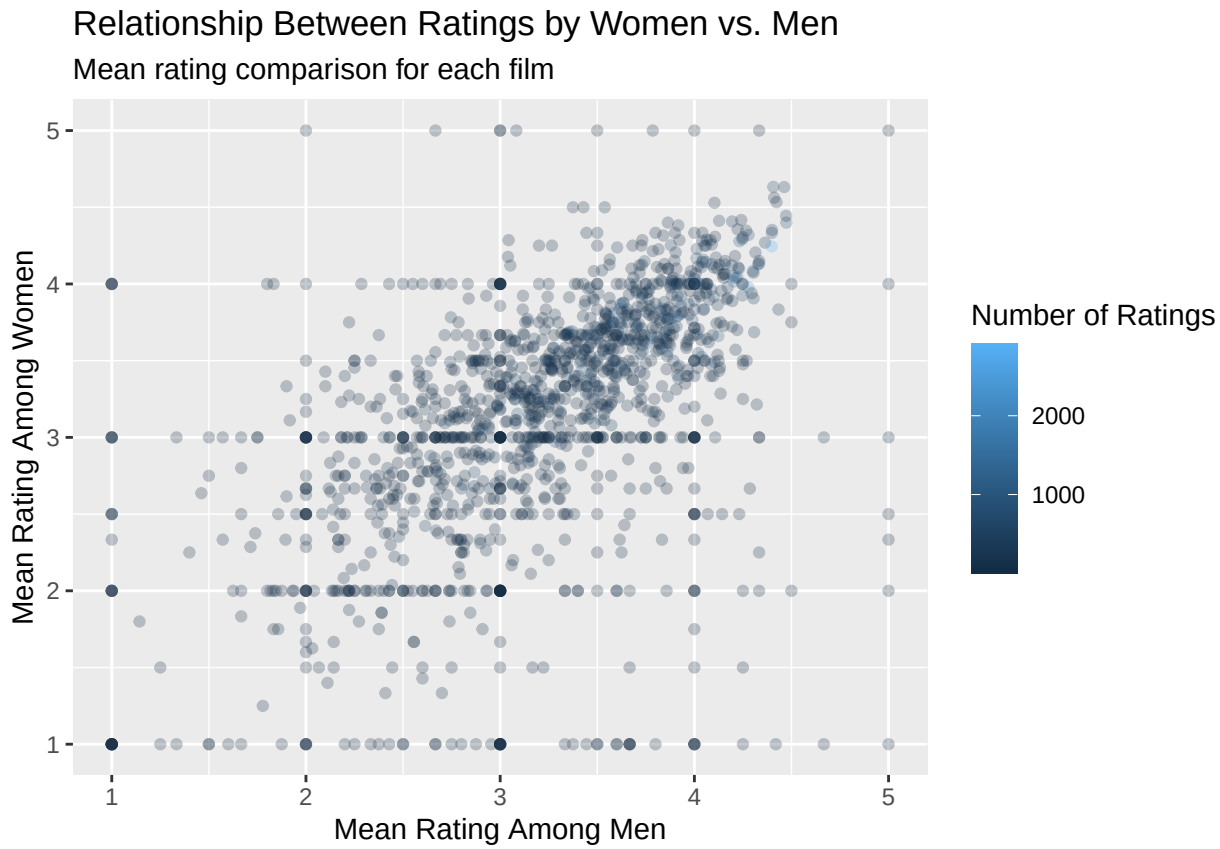
```
## [1] 3.078132
```

Problem 3: Correlation: Men versus women

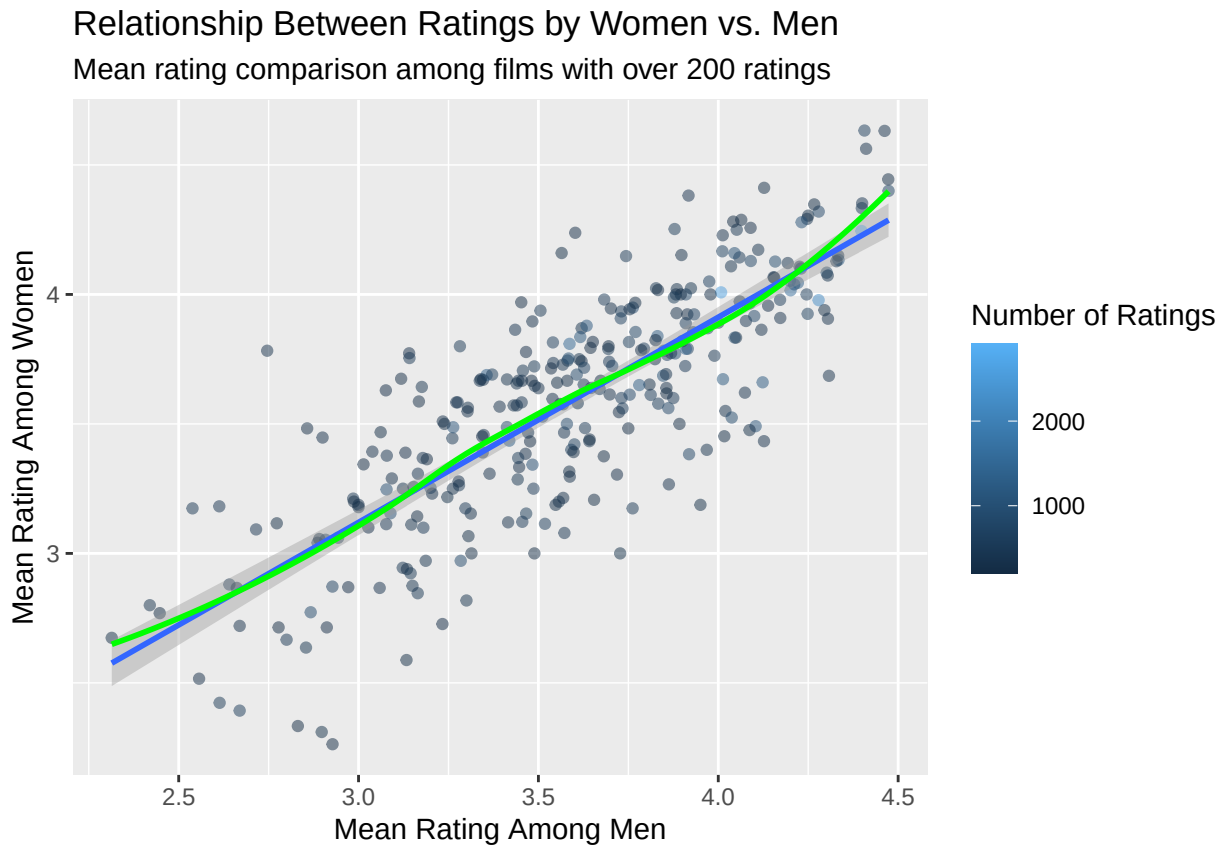
Let us look more closely at the relationship between the pieces of data we have.

- Make a scatter plot of men versus women and their mean rating for every movie.
- Make a scatter plot of men versus women and their mean rating for movies rated more than 200 times.
- Compute the correlation coefficent between the ratings of men and women.
  - When we compare mean ratings between men and women while including movies with 100 or fewer ratings, the correlation between mean rating among men and mean rating among women appears positive but not very strong for prediction. The correlation coefficient in this case is 0.5149489. When considering movies with more than 100 ratings the correlation is stronger with a correlation coefficient in this case of 0.8042434.
  - Considering movies with more than 100 ratings, the relationship between mean men rating and mean women rating is linear for the most part. This is more true near the mean of the mean ratings (about 3.5) where a rating of about 3.5 among men corresponds to a mean rating of about 3.5 among women. The relation appears not quite linear for high and low mean ratings.

```
## Warning: Removed 217 rows containing missing values (geom_point).
```




```
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



- Conjecture under what circumstances the rating given by one gender can be used to predict the rating given by the other gender.
  - Question: Are men and women more similar when they are younger or older?
  -

```
mlData_aggregates <- mlData %>%
  group_by(movie_title) %>%
  filter(gender == "M", age > 40) %>%
  summarise(mean_rating_men40plus = mean(rating)) %>%
  full_join(mlData_aggregates)
```

```
## Joining, by = "movie_title"
```

```
mlData_aggregates <- mlData %>%
  group_by(movie_title) %>%
  filter(gender == "W", age > 40) %>%
  summarise(mean_rating_women40plus = mean(rating)) %>%
  full_join(mlData_aggregates)
```

```
## Joining, by = "movie_title"
```

```
mlData_aggregates <- mlData %>%
  group_by(movie_title) %>%
  filter(gender == "M", age < 30) %>%
  summarise(mean_rating_men30minus = mean(rating)) %>%
  full_join(mlData_aggregates)
```

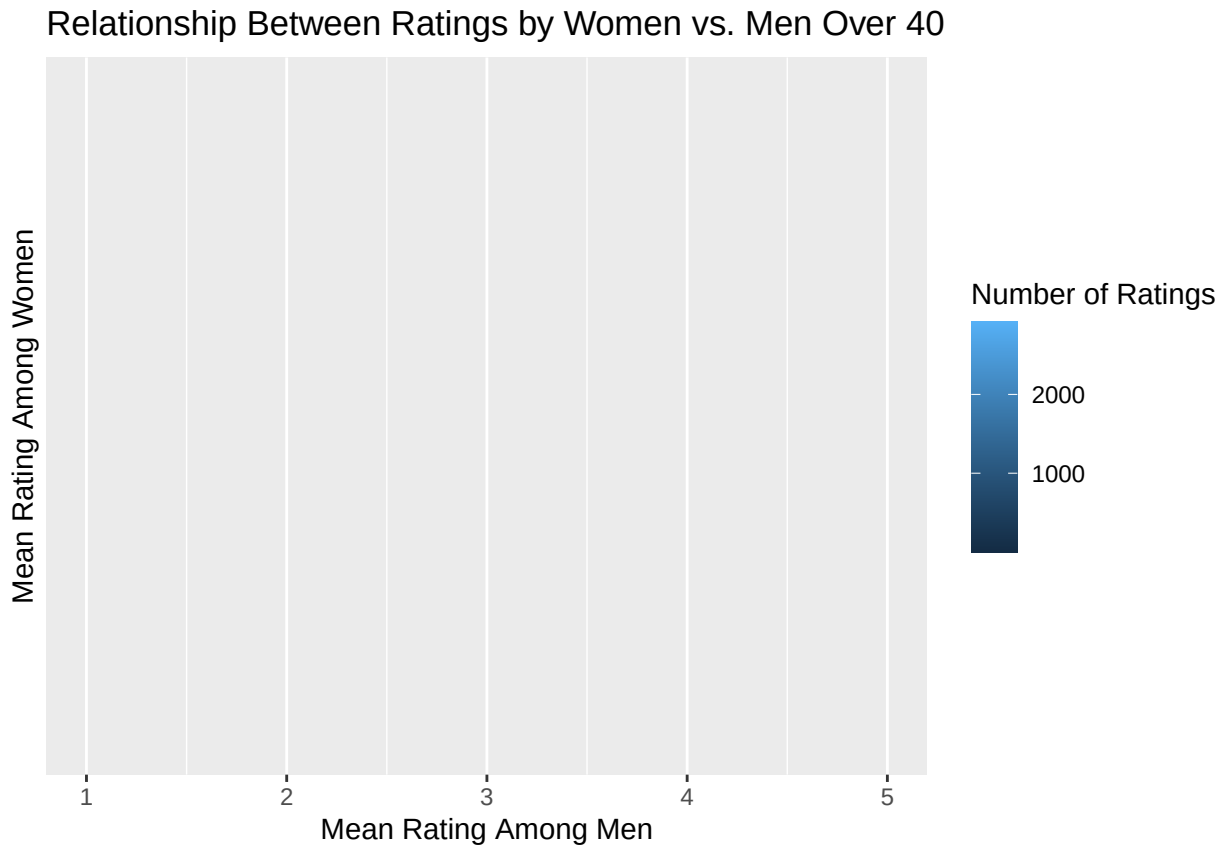
```
## Joining, by = "movie_title"
```

```
mlData_aggregates <- mlData %>%
  group_by(movie_title) %>%
  filter(gender == "W", age < 30) %>%
  summarise(mean_rating_women30minus = mean(rating)) %>%
  full_join(mlData_aggregates)
```

```
## Joining, by = "movie_title"
```

```
old_scatterplot <- mlData_aggregates %>%
  ggplot(aes(x = mean_rating_men40plus, y = mean_rating_women40plus)) +
  geom_point(aes(color = rating_count), alpha = .5) +
  labs(
    title = "Relationship Between Ratings by Women vs. Men Over 40",
    x = "Mean Rating Among Men",
    y = "Mean Rating Among Women",
    color = "Number of Ratings"
    )
old_scatterplot
```

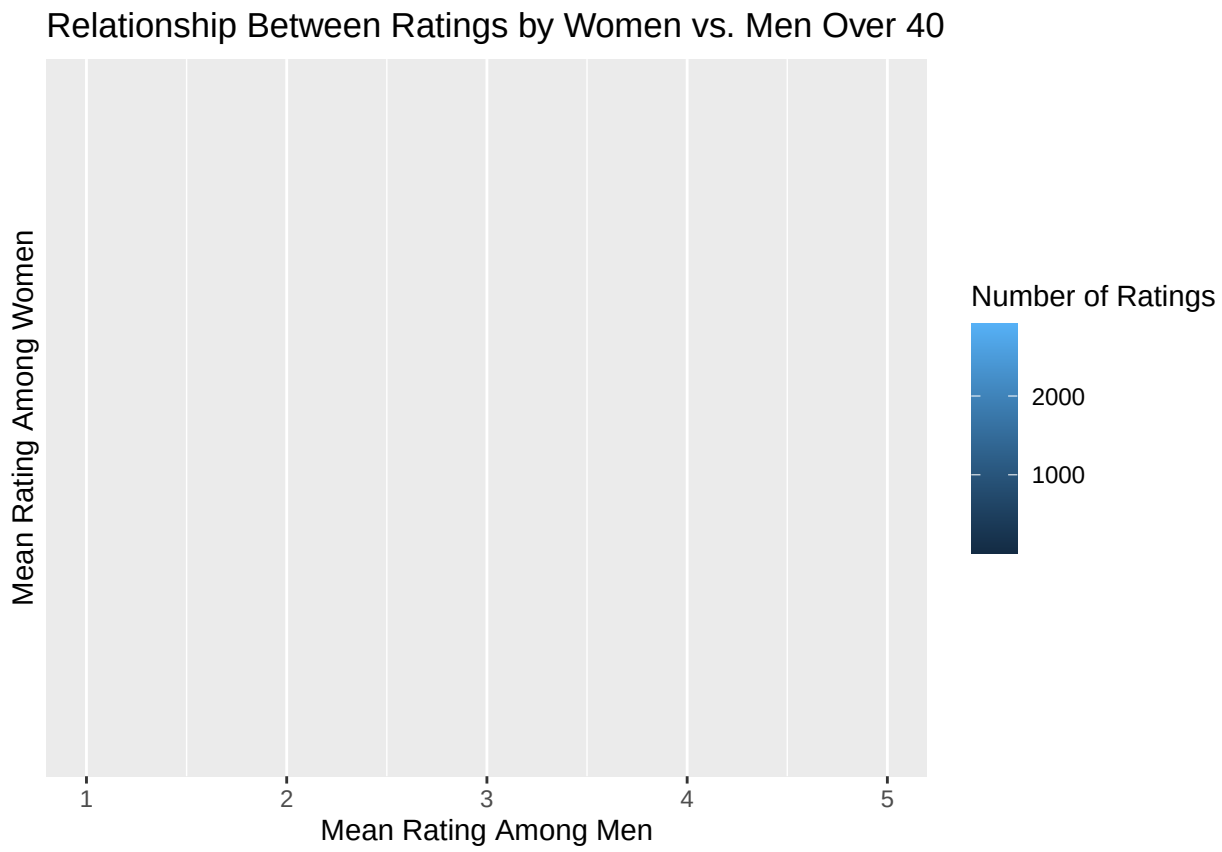
```
## Warning: Removed 1662 rows containing missing values (geom_point).
```




```
young_scatterplot <- mlData_aggregates %>%
  ggplot(aes(x = mean_rating_men30minus, y = mean_rating_women30minus)) +
  geom_point(aes(color = rating_count), alpha = .5) +
  labs(
    title = "Relationship Between Ratings by Women vs. Men Over 40",
    x = "Mean Rating Among Men",
```

```
    y = "Mean Rating Among Women",
    color = "Number of Ratings"
    )
young_scatterplot
```

```
## Warning: Removed 1662 rows containing missing values (geom_point).
```




```
gender_mean_corr_old = cor(
    mlData_aggregates$mean_rating_men40plus,
    mlData_aggregates$mean_rating_women40plus,
    )

gender_mean_corr_young = cor(
    mlData_aggregates$mean_rating_men30minus,
    mlData_aggregates$mean_rating_women30minus
    )
```

Problem 4: Open Ended Question: Business Intelligence

- From the exploration, I would suggest marketing films to lawyers, doctors, and educators of both genders. If we have discovered anything from this dataset, it is that men and women really do not differ significantly in their preferences and trying to make business decisions based on this factor is not recommended. Consider marketing Film-Noir and War Movies. If the film is released and not well received by the first few critics, elicit ratings from more critics. Generally, by increasing the number of ratings, the film is likely to improve it's overall mean and median ratings.